HELDA
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

UNIVERSITY OF HELSINKI

https://helda.helsinki.fi

In Silico Estimation of the Abundance and Phylogenetic Significance of the Composite Oct4-Sox2 Binding Motifs within a Wide Range of Species

Kulyyassov, Arman

2020-12

Kulyyassov , A & Kalendar , R 2020 , ' In Silico Estimation of the Abundance and Phylogenetic Significance of the Composite Oct4-Sox2 Binding Motifs within a Wide Range of Species ' , Data , vol. 5 , no. 4 , 111 . https://doi.org/10.3390/data5040111

http://hdl.handle.net/10138/323128
https://doi.org/10.3390/data5040111

cc_by_nd
publishedVersion

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

data

MDPI

Communication

In Silico Estimation of the Abundance and Phylogenetic Significance of the Composite Oct4-Sox2 Binding Motifs within a Wide Range of Species

Arman Kulyyassov [1,*] and Ruslan Kalendar [2,3,*]

[1] Republican State Enterprise "National Center for Biotechnology", 13/5 Kurgalzhynskoye Road, Nur-Sultan 010000, Kazakhstan
[2] Department of Agricultural Sciences, University of Helsinki, FI-00014 Helsinki, Finland
[3] National Laboratory Astana, Nazarbayev University, Nur-Sultan 010000, Kazakhstan
* Correspondence: kulyyasov@biocenter.kz (A.K.); ruslan.kalendar@helsinki.fi (R.K.)

Received: 5 October 2020; Accepted: 27 November 2020; Published: 29 November 2020

Abstract: High-throughput sequencing technologies have greatly accelerated the progress of genomics, transcriptomics, and metagenomics. Currently, a large amount of genomic data from various organisms is being generated, the volume of which is increasing every year. Therefore, the development of methods that allow the rapid search and analysis of DNA sequences is urgent. Here, we present a novel motif-based high-throughput sequence scoring method that generates genome information. We found and identified Utf1-like, Fgf4-like, and Hoxb1-like motifs, which are cis-regulatory elements for the pluripotency transcription factors Sox2 and Oct4 within the genomes of different eukaryotic organisms. The genome-wide analysis of these motifs was performed to understand the impact of their diversification on mammalian genome evolution. Utf1-like, Fgf4-like, and Hoxb1-like motif diversity was evaluated across genomes from multiple species.

Keywords: phylogeny; protein–protein interactions (PPI); in vivo DNA-dependent protein–protein interaction; pluripotency transcription factors Sox2 and Oct4; reprogramming

1. Introduction

The processes of cell reprogramming to a pluripotent state at the molecular level starts with protein–protein convergence caused by binding to neighboring DNA sites [1,2]. The transcription factors of pluripotency Sox2 (SRY-box 2), Oct4 (Octamer-binding transcription factor 4), and Nanog are key in the transcriptional network that controls stem cell pluripotency and the induction of pluripotency in somatic cells [3–5]. Sox2 belongs to a large group of Sox family proteins first discovered in 1990 [6–8]. All Sox proteins have the high-mobility group (HMG) box domain that may mediate non-sequence-specific and sequence-specific DNA binding [9]. HMG proteins are ubiquitous and abundant nuclear proteins that bind to nucleosomes and cause structural changes in chromatin. These non-histone proteins perform a significant role in DNA replication, recombination, transcription, and DNA repair processes. Most Sox TFs bind to and regulate different sets of genes in different cellular contexts. For example, Sox2 participates in a stunningly diverse class of cells and tissue types, including pluripotent stem cells, lung tissue, nerve lines, ear, and eye [10,11].

Among DNA-binding transcription factors, the POU genes represent a large group and play a fundamental role in cell-type specification and developmental regulation. The abbreviation POU originates from the names of three mammalian transcription factors, the pituitary-specific Pit-1, the octamer-binding proteins Oct-1 and Oct-2, and the neural Unc-86 from *Caenorhabditis elegans* [12]. Several detailed reviews have discussed DNA-binding by POU TFs and their function in mammalian

development [12–14]. Despite the fact that the domain is widespread across various species of living organisms, the sequence of this domain is nevertheless strictly conservative [15]. POU genes consist of the following three parts: A N-terminal POU-specific domain (POU'S'), a C-terminal homeodomain (POU'HD'), and a linker region between the two. OCT4 is a member of the octamer-binding subgroup of the POU family of transcription factors [16,17], which binds to the octamer motif (ATGCAAAT consensus sequence) using a bipartite DNA-binding POU domain.

It was previously described that Oct4 and Sox2 cooperatively control the pluripotent-specific expression of several genes by binding together with the cis-regulatory element Oct-Sox, and thereby regulate the transcription of important target genes, such as *Fgf4*, *Utf1*, *Pou5f1(Oct4)*, *Nanog*, and *Sox2* [11,18]. The binding sites of the DNA-associated pluripotency transcription factors were identified by the ChIPseq technique applied genome-wide for mapping TF binding regions in living cells [19,20]. The method combines chromatin immunoprecipitation (ChIP) using TF-specific antibodies with high-throughput next-generation parallel sequencing [21,22]. Along with this typical ChIP analysis, modern computational methods are also becoming increasingly important for genome-wide studies of protein–DNA interactions.

Nucleotide sequences outside of coding regions tend to be less conserved among organisms unless they are important for function, that is, where they are involved in the regulation of gene expression. Thus, the discovery of motifs in protein and nucleotide sequences can lead to the definition of function and clarification of the evolutionary relationship between sequences. In this work, we describe a novel approach, the DNA sequences profiling of human genome or other organisms, with the examples of Utf1-like, Fgf4-like and Hoxb1-like motifs, which are cis-regulatory elements for pluripotency transcription factors Sox2 and Oct4. We present validation results and provide examples that demonstrate its application in phylogenetic analyses between different species.

Genome-wide analyses of these motifs in a single species have been previously conducted to estimate common motifs, evolution, patterns of expression, and predicted localization. Multi-genome analyses of Utf1-like, Fgf4-like, and Hoxb1-like motifs resulted in a more rigorous and consistent description, and provided insight into their evolution. However, there is currently a lack of targeted studies investigating this question across multiple species. To aid in furthering our understanding of these regulatory elements, we performed multiple bioinformatics analyses to rigorously define the conserved motifs across multiple species, and to examine how they are spread throughout mammalian and other species.

2. Materials and Methods

2.1. Bioinformatic Identification of Sox2/Oct4 Motif Sequences

To analyze the frequencies of the Sox2/Oct4 motif sequences, we used FastPCR software (PrimerDigital, Helsinki, Finland) [23,24] with the "Restriction" tool that allowed custom degenerated sequences to search the genomes of various organisms. For this, the flag "-2" was used, which allows for collecting all cases of complementary coincidences both for one chromosome and in total for the entire genome. All other parameters were left at their default values. The program performed a forward and complementary search for each sequence motif. For each degenerated sequence motif, a table was generated indicating a specific sequence in the genome and its frequency.

Genomes for various representatives of mammals (*Homo sapiens*, *Sciurus vulgaris*, *Mus musculus*, *Rattus norvegicus*, *Capra hircus*, *Bos taurus*, *Oryctolagus cuniculus*, *Suss crofa*, *Felis catus*, *Canis lupus*, *Equus caballus*), birds (*Gallus gallus*), marsupials (*Sarcophilus harrisii*), plants (*Arabidopsis thaliana*, *Brachypodium distachyon*, *Sorghum bicolor*, *Solanum pennellii*, *Medicago truncatula*, *Glycine max*), insects (*Nasonia vitripennis*, *Apis mellifera*), amphibians (*Xenopus laevis*), and zebrafish (*Danio rerio*) obtained from the NCBI database of genome sequences (https://www.ncbi.nlm.nih.gov/genome/browse/) as a target set, and Sox2/Oct4 motifs as the query sets.

A list of the screened genomes that were searched is presented in Table S2 and S3.

The use of Sox2/Oct4 motif sequences allowed the assessment of how the software handles complex tasks, such as in silico degenerate pattern searching. The sequence motif lengths can be at least 4 nucleotides. The analysis results include a table and are presented separately for each sequence motif and its frequency. The program "decodes" the degenerated sequence motif and presents the frequencies for each of them. For each genome under study, tables were obtained in a tab-delimited format, and were combined and sorted by text.

Additionally, neutral sequences for control analysis used random sequences, for example, inverted Utf1-like sequence (ATGYWDGDnHWTTSW).

The resulting Sox2/Oct4 motif sequence was used as the search query with FastPCR software against all studied genome sequences (listed in Table S2) using the flag "-2" with all other settings at their default values. This resulting sequence dataset (Table S3) was used for all subsequent analyses.

2.2. A Phylogenetic Tree of Sox2/Oct4 Motif Sequences

The Sox2/Oct4 motifs discovered by FastPCR software [23,24] were used as input to define the phylogeny of studied genomes. To determine the number of sites and the average distance between them, an in-house script was developed to analyze the FastPCR results. Likewise, an in-house script was used to compute the Nei's standard genetic distance [25] between each species based on a comparison of the frequencies of each motif between species:

$$D = -ln\frac{J_{xy}}{\sqrt{J_x\, J_y}}$$

where J_{xy} is calculated as the sum of shared frequencies of a motif sequence by genome x and y normalized for each genome. Large and small genomes will have different values for each motif; therefore, the compared frequencies of the motifs were calculated for each specific genome. The J_{xy} value ranges from 0 to 1, depending on the degree of frequency coincidence between two unrelated genomes. The J_x or J_y values correspond to the number of motifs for genome x and y.

A phylogenetic tree of all the species used in this study was created using the MEGA X software (Pennsylvania State University, USA) [26].

3. Results and Discussion

Based on results from the proximity utilizing the biotinylation (PUB) method in a living cell, we previously discovered a high level of biotin labeling of transcription factors Sox2 and Oct4 in comparison with various control proteins [27]. The Sox2 protein binds to the CATTGTT sequence, while Oct4 recognizes the octameric consensus ATGCTAGT sequence [18,28–39], such as in undifferentiated embryonic cell transcription factor 1 (Utf1) or other motifs (Supplementary material Figure S1). Usually, these recognizing DNA sites are linked together to form a composite motif, known also as the "canonical" motif. Thus, the interaction between these key transcription factors of pluripotency proceeds via the DNA-binding domains HMG, POU'S', and POU'HD'.

An analysis of the literature on the distribution of motif variants in genomes showed the presence of some differences in the sequences. For example, in the mouse genome, there are two variants of the motifs Fgf4 (1 and 2 points of Table 1) and Utf1 (6 and 7 points of Table 1). For both motifs, matches were found in the genomic database Ensembl (Supplementary Table S2). It is interesting to note that other sequences in which there are combinatorial replacements of the Oct4 moiety in the Utf1 motif ATGCTAGT with another ATGCTAGA are also present in the mouse genome. Therefore, we searched for DNA sequences using generalized formulas.

Table 1. Distribution of different variants of canonical motifs in the mouse genome (*Mus musculus*, C57BL/6NJ). Data in the last column were obtained using Ensembl Genome Browser.

Motif	Sequence	Reference	Number of Hits Found on Ensembl
	SoxOct		
Fgf4	CTTTGTTTGAATGCTAAT	[11]	32
Fgf4	CTTTGTTTGGATGCTAAT	[30,37,40–42]	37
Pou5f1	CTTTGTTATGCATCT	[11,40,41]	12
Sox2	CATTGTGATGCATAT	[11,40,41,43]	8
Nanog	CATTGTAATGCAAAA	[11,40,41,44]	13
Utf1	CATTGTTATGCTAGT	[11,29,30,40,41,43,45]	4
Utf1	CATTGTTATGCTAGA	[18]	1
HoxB1	CTTTGTCATGCTAAT	[18]	14
Fbxo15	CATTGTTATGATAAA	[11,41,46]	32
Dppa4	ATTTGTAaATGCTAAA	[11]	47
Gsh2	CTTTGTCATGCAGAG	[18]	16
Nes	ATGCTAATtattgccTTTTGTC	[11]	62

To generate variants of the Sox2/Oct4 motifs, we used the FastPCR software [23,24] with the "Restriction" tool that allowed custom degenerated sequences to search in the genomes of various organisms. According to the IUPAC nucleotide nomenclature rules, the following generalized motif sequences as described in the manuscript of Tapia et al. [28] were used: Fgf4-like (HWTTSWnnnnATGYWDWD), Utf1-like (HWTTSWnATGYWDGD), and Hoxb1-like (HWTTSWnATGYWDWD). The largest numbers of sites were found for the Fgf4-like and Hoxb1-like motifs (Supplementary material Table S1).

Determining the molecular genetic relationships of organisms can be performed on the common features for all organisms. As a rule, for this purpose, universal genes characteristics of living organisms are chosen, such as ribosomal RNA genes.

In this study, we suggested that for genetically related species, there should also be a general trend for evolutionarily ancient promoter regions, such as Utf1-like, Fgf4-like, and Hoxb1-like canonic variant motifs. To test this, we analyzed the genome-wide sequences of some closely related and distant mammalian species, and some representatives of marsupials, birds, amphibians, insects, and plants. Since the studied sequences are characterized as promoter regions only for mammals, the genetic relationship should be well traced precisely among animal species. In the plant kingdom and for insects, these sequences have a different evolutionary nature.

To determine the distance coefficient between the compared species according to Utf1-like, Fgf4-like, and Hoxb1-like canonic variant motifs, we proceeded from the following. For related species, the amount of each competitive sequence must be similar given the size of the genome. The larger the genome size, the more variants of a particular sequence will be observed for each canonic variant motif. The closer the genomes are, the more similarity there will be in the frequency and sequence quality for each canonic motif. The total distance is defined as the sum of all coefficients for all frequencies of each sequence in each canonic motif. To calculate such distances, we used Nei's standard genetic distance, considering the coincidence of individual sites and their frequencies, and normalized to the total number of sites per genome for the species being compared.

Next, we examined the distribution of Utf1-like, Fgf4-like, and Hoxb1-like canonic variant motifs for various members of the animal kingdom.

For the Hoxb1-like canonic variant motif (HWTTSWnATGYWDWD), a maximum of 107,723 sites were identified in the *Mus musculus* genome and a minimum of 4304 for the insect *Nasonia vitripennis*. For the *Sciurus vulgaris* and *Mus musculus* genomes, the largest number of variants of this motif was identified (6910 unique sequences). The plant genomes were also characterized by a low number

of sites on the Hoxb1-like canonic variant motif genome. The genome of *Gallus gallus* occupied an intermediate position between plants and mammals. The number of Hoxb1-like canonic variant motifs directly depended on the genome size. In the phylogenetic analysis, we clearly observed the tendency of species distribution according to their phylogeny (Figure 1).

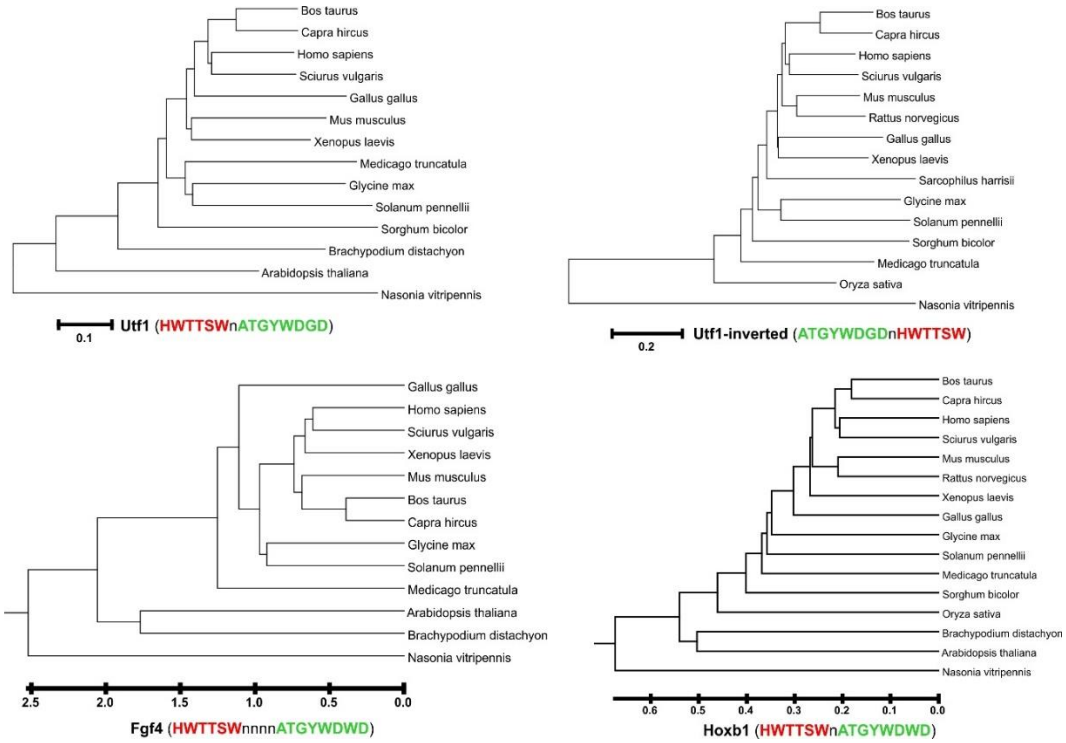

Figure 1. Dendrograms within species using Utf1-like and Utf1-inverted canonic variant motifs calculated by the Minimum Evolution (ME) and UPGMA methods for Fgf4-like and Hoxb1-like canonic variant motifs. The sequences of the motifs are represented in the standard IUB/IUPAC nucleic acid codes (http://blast.ncbi.nlm.nih.gov/blast/fasta.shtml).

The frequency trend of these sites is in good agreement with the relatedness of these species. For very closely related species, these values practically coincide.

An analysis of the occurrence frequency of the Utf1-like canonic variant motif (HWTTSWnATGYWDGD) site for the studied species showed that the minimum number of this site in the insect *Nasonia vitripennis* is 1077, and the maximum in humans is 83,559. For the *Sciurus vulgaris* genome, the largest number of variants of this motif was revealed (3451 unique sequences for the human genome). The phylogenetic analysis for this site generally coincides with what we observed already for the Hoxb1-like site, with some changes regarding the position of *Gallus gallus* among mammals. This is an interesting fact that is associated with the high number of unique sequences of this motif in the genome of *Gallus gallus* (3287), with a low number of these sites per genome (16,515).

An analysis of the frequency of occurrence of the site Fgf4-like canonic variant motif (HWTTSWnnnnATGYWDWD) revealed the largest variants of this motif (192,940 sites and 114,914 unique variants for the human genome). The minimum number of this site in the insect *Nasonia vitripennis* was 4507, with 4317 unique variants. Phylogenetic analysis was performed by the UPGMA method, which most accurately characterizes the genetic relationship of the studied species. In general, the phylogenetic analysis for this site overall coincides with what we already observed for the Hoxb1-like and Utf1-like sites; some changes concerning the position of *Gallus gallus* were isolated from mammals' genomes, which better corresponds to the phylogeny of these species.

To examine the distribution of the Sox2/Oct4 motif sequences among mammals in more detail, we made a simplified grouping of the studied species. Because these clades consist of a different

number of species, all of the population fractions were reweighted to normalize the results so as to facilitate comparison.

4. Conclusions

In this work, we performed a genome-wide analysis of the cis-regulatory elements and promoters for several mammals and distantly related species, including animals such as insects and amphibians and some plant species. We also studied the extent to which these data are common among these eukaryotes in accordance with genome size and evolutionary relationship between these species. Our hypothesis was that closely related species should have a similar frequency of occurrence of these sequences in relation to the genome size. If these sequences are evolutionarily neutral, then we can trace them in a variety of species that are evolutionarily distant from mammals. In the case of the evolutionary significance of these sequences, we could observe the absence of any connection between the frequencies of these sequences and the phylogeny of the compared species. To perform the analysis, we wrote a script and used FastPCR programs to search for these sequences genome-wide. We selected a mathematical apparatus for calculating genetic distances for the data used (a list of sequences, and their frequencies in relation to the size of the genome).

The results of this work indicate that the genome-wide analysis of the frequencies of cis-regulatory elements is shown to be neutral. That is, we can apply these data for phylogenetic analysis. Thus, we have shown that our assumption in the analysis of the frequencies of certain sequences, including sequences of regulatory elements and promoters, can be used to calculate evolutionary distances and construct a phylogenetic tree. Although the investigated sequences of regulatory elements and promoters have a functional role and should be subjected to selection, we observed that these sequences are evolutionarily neutral and applicable for revealing the relationship of the compared genomes.

Supplementary Materials: Supplementary Materials can be found at http://www.mdpi.com/2306-5729/5/4/111/s1. Figure S1: Detection of protein-protein interaction with biotin ligase and a biotin acceptor peptide substrate; Table S1: Distribution of Utf1-like, Fgf4-like and Hoxb1-like canonic variant motifs in human genome; Table S2: Distribution of canonical motifs in the mouse genome (*Mus musculus*, C57BL/6NJ), obtained using Ensembl Genome Browser; Table S3: Distribution of Utf1-like, Fgf4-like, and Hoxb1-like motifs in the genomes for various representatives of mammals, birds, marsupials, plants, insects, amphibians, and zebrafish obtained with FastPCR software.

Author Contributions: A.K. collected and analyzed data, wrote the manuscript, managed the project, and acquired financial support for the project leading to this publication; R.K. analyzed data using FastPCR software, wrote the manuscript, and provided critical reviews and commentary to the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a grant from the Ministry of education and science of the Republic of Kazakhstan (AP05132131). Open access funding was provided by University of Helsinki including Helsinki University Central Hospital. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments: We warmly thank Alexander Bolshoy (Palacký University Olomouc, Czech Republic) for critical reading of the manuscript.

Conflicts of Interest: The authors declare that they have no conflict of interest.

Abbreviations

BAP	Biotin Acceptor Peptide
BirA	Biotinylating ligase and repressor of biotin biosynthesis of *Escherichia coli*
ChIP	Chromatin immunoprecipitation
ESCs	Embryonic stem cells
Fgf4	Fibroblast growth factor 4
HMG	High-mobility group
Hox-B1	Homeobox protein
iPSCs	Induced pluripotent stem cells
NMR	Nuclear magnetic resonance spectroscopy
Oct4	Octamer-binding transcription factor 4

POU'HD'	POU homeodomain
POU'S'	POU-specific domain
PPI	Protein-protein interactions
PUB	Proximity Utilizing Biotinylation
Sox2	SRY-box 2
TF	Transcription factors
Utf1	Undifferentiated embryonic cell transcription factor 1

References

1. Takahashi, K.; Yamanaka, S. A decade of transcription factor-mediated reprogramming to pluripotency. *Nat. Rev. Mol. Cell Biol.* **2016**, *17*, 183–193. [CrossRef] [PubMed]
2. Yu, J.; Vodyanik, M.A.; Smuga-Otto, K.; Antosiewicz-Bourget, J.; Frane, J.L.; Tian, S.; Nie, J.; Jonsdottir, G.A.; Ruotti, V.; Stewart, R.; et al. Induced pluripotent stem cell lines derived from human somatic cells. *Science* **2007**, *318*, 1917–1920. [CrossRef] [PubMed]
3. Takahashi, K.; Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **2006**, *126*, 663–676. [CrossRef] [PubMed]
4. Takahashi, K.; Tanabe, K.; Ohnuki, M.; Narita, M.; Ichisaka, T.; Tomoda, K.; Yamanaka, S. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **2007**, *131*, 861–872. [CrossRef] [PubMed]
5. Li, M.; Belmonte, J.C. Ground rules of the pluripotency gene regulatory network. *Nat. Rev. Genet.* **2017**, *18*, 180–191. [CrossRef] [PubMed]
6. Gubbay, J.; Collignon, J.; Koopman, P.; Capel, B.; Economou, A.; Munsterberg, A.; Vivian, N.; Goodfellow, P.; Lovellbadge, R. A gene-mapping to the sex-determining region of the mouse y-chromosome is a member of a novel family of embryonically expressed genes. *Nature* **1990**, *346*, 245–250. [CrossRef]
7. Sinclair, A.H.; Berta, P.; Palmer, M.S.; Hawkins, J.R.; Griffiths, B.L.; Smith, M.J.; Foster, J.W.; Frischauf, A.M.; Lovell-Badge, R.; Goodfellow, P.N. A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature* **1990**, *346*, 240–244. [CrossRef]
8. Bowles, J.; Schepers, G.; Koopman, P. Phylogeny of the sox family of developmental transcription factors based on sequence and structural indicators. *Dev. Biol.* **2000**, *227*, 239–255. [CrossRef]
9. Stros, M.; Launholt, D.; Grasser, K.D. The hmg-box: A versatile protein domain occurring in a wide variety of DNA-binding proteins. *Cell Mol. Life Sci.* **2007**, *64*, 2590–2606. [CrossRef]
10. Yesudhas, D.; Batool, M.; Anwar, M.A.; Panneerselvam, S.; Choi, S. Proteins recognizing DNA: Structural uniqueness and versatility of DNA-binding domains in stem cell transcription factors. *Genes* **2017**, *8*, 192. [CrossRef]
11. Kamachi, Y.; Kondoh, H. Sox proteins: Regulators of cell fate specification and differentiation. *Development* **2013**, *140*, 4129–4144. [CrossRef] [PubMed]
12. Phillips, K.; Luisi, B. The virtuoso of versatility: Pou proteins that flex to fit. *J. Mol. Biol.* **2000**, *302*, 1023–1039. [CrossRef] [PubMed]
13. Tantin, D. Oct transcription factors in development and stem cells: Insights and mechanisms. *Development* **2013**, *140*, 2857–2866. [CrossRef] [PubMed]
14. Jerabek, S.; Merino, F.; Scholer, H.R.; Cojocaru, V. Oct4: Dynamic DNA binding pioneers stem cell pluripotency. *Biochim. Biophys. Acta* **2014**, *1839*, 138–154. [CrossRef]
15. Gold, D.A.; Gates, R.D.; Jacobs, D.K. The early expansion and evolutionary dynamics of pou class genes. *Mol. Biol. Evol.* **2014**, *31*, 3136–3147. [CrossRef]
16. Malik, V.; Zimmer, D.; Jauch, R. Diversity among pou transcription factors in chromatin recognition and cell fate reprogramming. *Cell Mol. Life Sci.* **2018**, *75*, 1587–1612. [CrossRef]
17. Esch, D.; Vahokoski, J.; Groves, M.R.; Pogenberg, V.; Cojocaru, V.; Vom Bruch, H.; Han, D.; Drexler, H.C.; Arauzo-Bravo, M.J.; Ng, C.K.; et al. A unique oct4 interface is crucial for reprogramming to pluripotency. *Nat. Cell Biol.* **2013**, *15*, 295–301. [CrossRef]
18. Jauch, R.; Aksoy, I.; Hutchins, A.P.; Ng, C.K.; Tian, X.F.; Chen, J.; Palasingam, P.; Robson, P.; Stanton, L.W.; Kolatkar, P.R. Conversion of sox17 into a pluripotency reprogramming factor by reengineering its association with oct4 on DNA. *Stem Cells* **2011**, *29*, 940–951. [CrossRef]

19. Lloyd, S.M.; Bao, X. Pinpointing the genomic localizations of chromatin-associated proteins: The yesterday, today, and tomorrow of chip-seq. *Curr. Protoc. Cell Biol.* **2019**, *84*, e89. [CrossRef]
20. Lai, X.; Stigliani, A.; Vachon, G.; Carles, C.; Smaczniak, C.; Zubieta, C.; Kaufmann, K.; Parcy, F. Building transcription factor binding site models to understand gene regulation in plants. *Mol. Plant* **2019**, *12*, 743–763. [CrossRef]
21. Robertson, G.; Hirst, M.; Bainbridge, M.; Bilenky, M.; Zhao, Y.; Zeng, T.; Euskirchen, G.; Bernier, B.; Varhol, R.; Delaney, A.; et al. Genome-wide profiles of stat1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **2007**, *4*, 651–657. [CrossRef] [PubMed]
22. Valouev, A.; Johnson, D.S.; Sundquist, A.; Medina, C.; Anton, E.; Batzoglou, S.; Myers, R.M.; Sidow, A. Genome-wide analysis of transcription factor binding sites based on chip-seq data. *Nat. Methods* **2008**, *5*, 829–834. [CrossRef]
23. Kalendar, R.; Khassenov, B.; Ramankulov, Y.; Samuilova, O.; Ivanov, K.I. Fastpcr: An in silico tool for fast primer and probe design and advanced sequence analysis. *Genomics* **2017**, *109*, 312–319. [CrossRef] [PubMed]
24. Kalendar, R.; Lee, D.; Schulman, A.H. Java web tools for pcr, in silico pcr, and oligonucleotide assembly and analysis. *Genomics* **2011**, *98*, 137–144. [CrossRef]
25. Nei, M. Genetic distance between populations. *Am. Nat.* **1972**, *106*, 283–292. [CrossRef]
26. Stecher, G.; Tamura, K.; Kumar, S. Molecular evolutionary genetics analysis (mega) for macos. *Mol. Biol. Evol.* **2020**, *37*, 1237–1239. [CrossRef] [PubMed]
27. Kulyyassov, A.; Ogryzko, V. In vivo quantitative estimation of DNA-dependent interaction of sox2 and oct4 using bira-catalyzed site-specific biotinylation. *Biomolecules* **2020**, *10*, 142. [CrossRef]
28. Tapia, N.; MacCarthy, C.; Esch, D.; Gabriele Marthaler, A.; Tiemann, U.; Arauzo-Bravo, M.J.; Jauch, R.; Cojocaru, V.; Scholer, H.R. Dissecting the role of distinct oct4-sox2 heterodimer configurations in pluripotency. *Sci. Rep.* **2015**, *5*, 13533. [CrossRef]
29. Scholer, H.R.; Ruppert, S.; Suzuki, N.; Chowdhury, K.; Gruss, P. New type of pou domain in germ line-specific protein oct-4. *Nature* **1990**, *344*, 435–439. [CrossRef]
30. Remenyi, A.; Lins, K.; Nissen, L.J.; Reinbold, R.; Scholer, H.R.; Wilmanns, M. Crystal structure of a pou/hmg/DNA ternary complex suggests differential assembly of oct4 and sox2 on two enhancers. *Genes Dev.* **2003**, *17*, 2048–2059. [CrossRef]
31. Merino, F.; Ng, C.K.L.; Veerapandian, V.; Scholer, H.R.; Jauch, R.; Cojocaru, V. Structural basis for the sox-dependent genomic redistribution of oct4 in stem cell differentiation. *Structure* **2014**, *22*, 1274–1286. [CrossRef] [PubMed]
32. Hou, L.; Srivastava, Y.; Jauch, R. Molecular basis for the genome engagement by sox proteins. *Semin Cell Dev. Biol.* **2017**, *63*, 2–12. [CrossRef] [PubMed]
33. Williams, D.C., Jr.; Cai, M.; Clore, G.M. Molecular basis for synergistic transcriptional activation by oct1 and sox2 revealed from the solution structure of the 42-kda oct1.Sox2.Hoxb1-DNA ternary transcription factor complex. *J. Biol. Chem.* **2004**, *279*, 1449–1457. [CrossRef] [PubMed]
34. Parslow, T.G.; Blair, D.L.; Murphy, W.J.; Granner, D.K. Structure of the 5′ ends of immunoglobulin genes—A novel conserved sequence. *Proc. Natl. Acad. Sci. USA* **1984**, *81*, 2650–2654. [CrossRef] [PubMed]
35. Boyer, L.A.; Lee, T.I.; Cole, M.F.; Johnstone, S.E.; Levine, S.S.; Zucker, J.P.; Guenther, M.G.; Kumar, R.M.; Murray, H.L.; Jenner, R.G.; et al. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **2005**, *122*, 947–956. [CrossRef]
36. Shlyueva, D.; Stampfel, G.; Stark, A. Transcriptional enhancers: From properties to genome-wide predictions. *Nat. Rev. Genet.* **2014**, *15*, 272–286. [CrossRef]
37. Kamachi, Y.; Uchikawa, M.; Tanouchi, A.; Sekido, R.; Kondoh, H. Pax6 and sox2 form a co-DNA-binding partner complex that regulates initiation of lens development. *Genes Dev.* **2001**, *15*, 1272–1286. [CrossRef]
38. Ambrosetti, D.C.; Basilico, C.; Dailey, L. Synergistic activation of the fibroblast growth factor 4 enhancer by sox2 and oct-3 depends on protein-protein interactions facilitated by a specific spatial arrangement of factor binding sites. *Mol. Cell. Biol.* **1997**, *17*, 6321–6329. [CrossRef]
39. Yuan, H.B.; Corbi, N.; Basilico, C.; Dailey, L. Developmental-specific activity of the fgf-4 enhancer requires the synergistic action of sox2 and oct-3. *Gene Dev.* **1995**, *9*, 2635–2645. [CrossRef]
40. Mistri, T.K.; Arindrarto, W.; Ng, W.P.; Wang, C.; Lim, L.H.; Sun, L.; Chambers, I.; Wohland, T.; Robson, P. Dynamic changes in sox2 spatio-temporal expression promote the second cell fate decision through fgf4/fgfr2 signaling in preimplantation mouse embryos. *Biochem. J.* **2018**, *475*, 1075–1089. [CrossRef]

41. Chew, J.L.; Loh, Y.H.; Zhang, W.; Chen, X.; Tam, W.L.; Yeap, L.S.; Li, P.; Ang, Y.S.; Lim, B.; Robson, P.; et al. Reciprocal transcriptional regulation of pou5f1 and sox2 via the oct4/sox2 complex in embryonic stem cells. *Mol. Cell Biol.* **2005**, *25*, 6031–6046. [CrossRef] [PubMed]
42. Ambrosetti, D.C.; Scholer, H.R.; Dailey, L.; Basilico, C. Modulation of the activity of multiple transcriptional activation domains by the DNA binding domains mediates the synergistic action of sox2 and oct-3 on the fibroblast growth factor-4 enhancer. *J. Biol. Chem.* **2000**, *275*, 23387–23397. [CrossRef] [PubMed]
43. Kamachi, Y.; Uchikawa, M.; Kondoh, H. Pairing sox off: With partners in the regulation of embryonic development. *Trends Genet.* **2000**, *16*, 182–187. [CrossRef]
44. Jung, M.; Peterson, H.; Chavez, L.; Kahlem, P.; Lehrach, H.; Vilo, J.; Adjaye, J. A data integration approach to mapping oct4 gene regulatory networks operative in embryonic stem cells and embryonal carcinoma cells. *PLoS ONE* **2010**, *5*, e10709. [CrossRef]
45. Nishimoto, M.; Miyagi, S.; Katayanagi, T.; Tomioka, M.; Muramatsu, M.; Okuda, A. The embryonic octamer factor 3/4 displays distinct DNA binding specificity from those of other octamer factors. *Biochem. Biophys. Res. Commun.* **2003**, *302*, 581–586. [CrossRef]
46. Tokuzawa, Y.; Kaiho, E.; Maruyama, M.; Takahashi, K.; Mitsui, K.; Maeda, M.; Niwa, H.; Yamanaka, S. Fbx15 is a novel target of oct3/4 but is dispensable for embryonic stem cell self-renewal and mouse development. *Mol. Cell Biol.* **2003**, *23*, 2699–2708. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).